An automated record linkage and pseudonymisation tool to support the efficient use of electronic health record data in clinical trials

Abstract

Introduction: Clinical trials are increasingly making use of routinely collected electronic health record data to answer clinical outcomes. However, making clinical data available to researchers in a useful format, in line with patient privacy and data protection requirements, presents a significant technical challenge. This project aimed to design and build a data linkage and pseudonymisation tool which can be used by sites to process data for research. An accompanying template data flow diagram will seek to clarify the data flows involved in a complex clinical trial.

Methods: A template Data Linkage and Pseudonymisation Tool (DLPT) was developed and adapted to the needs of an active cluster-randomised controlled trial with complex data flows including patient and staff record linkage. Accompanying documents were produced to provide guidance on adapting the DLPT to other trials, and a template data flow diagram was constructed.

Results: A DLPT adapted to the TYPPEX RCT was successfully implemented in trial and used by a participating site to process and export data in the format required by the research team. This represents a considerable time saving for the clinical trials unit, of 4-5 person hours per site per month.

Dissemination: The DLPT, accompanying templates and guidance document, and the template data flow diagram will be made available via the Norwich CTU website. These resources will be advertised more widely to CTUs via the UKCRC newsletter.

Introduction

Traditional patient identification and recruitment processes requiring patients to be identified, contacted and consented by their clinician are expensive and difficult to implement due to staff workload and limited time for research. This is especially problematic in primary care settings and community mental health services.[1] Furthermore, the use of informed consent as a mechanism for trial participation can be hampered by selection bias and poor recruitment.[2]

Alternative approaches to both identification of potentially eligible patients and collection of baseline and outcome data for trials have made use of Electronic Health Records (EHRs). To maximise patient participation, reduce selection bias, and increase efficiency, researchers are increasingly asking sites to provide routine data as part of the trial dataset, including in some cases in a de-identified format for eligible patients who are not approached to provide informed consent. However, there is huge variability in NHS staff time and skills in extracting, processing, and providing this data to researchers, and there are few researchers with the skills, experience and time to support NHS staff in providing this data. The ability to link

datasets, matching EHR data with trial research data such as patient-reported outcome measures, is crucial to trial analysis and represents a technical challenge for trialists.

The TYPPEX programme (NIHR PGfAR, RP-PG-0616-20003) includes a three-year stepped-wedge cluster randomised controlled trial with complex data collection and linkage.[3] The monthly record matching and data processing required for this trial is time-consuming and inefficient, with multiple transfers of patient data between Norwich CTU and sites. Our involvement in this trial prompted us to explore the development of an efficient and automated record matching and linkage tool which could be adapted for use in a range of trial designs and with a wide range of data sources.

In our experience, the details of data flows can be easily overlooked in early stages of trial development, leading to implementation problems and delays later on. We propose that a well-designed data flow diagram (DFD) for a complex trial would also be a useful output of the project, and could be used in the following scenarios:

- Trial funding application development and funder review
- Protocol review by NHS or institutional research ethics committees
- Patient and Public Involvement (PPI) review of trial design and patient-facing information
- Trial project management and oversight by the trial manager and oversight committees
- Applications for access to national health datasets (e.g. data held by NHS Digital)
- Applications to the Confidentiality Advisory Group for 'section 251' approval for use of data without patient consent

In this report we describe the development of the template data linkage and pseudonymisation tool (DLPT) and the adaptation of the tool to the TYPPEX trial. The DLPT package includes the template DLPT, guidance on the set-up of the DLPT for researchers/data managers, template instructions for site users and a template data flow diagram.

Methods

1. Data Linkage and Pseudonymisation Tool

A template data linkage and pseudonymisation tool was constructed to perform two main functions:

I. Match clinical patient records with those already held by researchers in a trial database
II. For unmatched patients assign a new study ID and fully de-identify the data prior to transfer to the researchers.

The tool was primarily designed for trials using a REDCap database,[4] which is the software used by Norwich CTU and many other academic institutions. However, where possible, alternative approaches and workarounds are suggested to ensure that researchers using a different software system can still use the DLPT.

The DLPT is a macro-enabled Excel workbook (.xlsm), with the data processing coded using Visual Basic for Applications (VBA). This programming language is well established, and .xlsm files have the advantage of being acceptable to the IT systems of most NHS organisations without (or with minimal) additional approvals, installation, or cost.

The set-up of the DLPT for a trial is described in detail in the accompanying DLPT Implementation Guide for Researchers and Data Managers. In summary, the researcher should adapt the DLPT to a trial by taking the following points into consideration:

- **Source dataset**: the data specification for the source data must be defined, and field names, and the order of columns in the data known. Typically, a custom-built report in the source database is constructed either by the site user or their software provider. The specification for this should be agreed by all parties and a blank report sent to the researcher showing the format and column headers that will be exported.

- **Trial database**: if the data is to be directly imported into a trial database, the database must contain the relevant fields, and the 'Processed data' sheet in the DLPT should be set up to export data in the format required for import into the trial database.

- **Link identifier**: The DLPT uses a single identifier to link patient records. Researchers should consider the most appropriate identifier for this purpose. Typically, NHS number is used.

- **Study ID:** The DLPT is designed to process data for multi-site clinical trials, where the study ID incorporates a site identifier as well as the participant number. E.g. *SP01-001* where *SP01* = site and *001* = participant. The format of the study ID must be defined in the DLPT during set-up.

We planned to demonstrate the utility of the DLPT in the TYPPEX RCT. The trial requires routinely collected clinical data from the Improving Access to Psychological Therapy (IAPT) services of three NHS mental health trusts for the primary outcome (patient recovery).

Patient data is recorded in our participating IAPT services using two software systems: iaptus (Mayden, Bath, UK) and PCMIS (University of York, UK). Once the data specification for the trial had been finalised, both software providers were asked to build a custom report to export trial data, for use by data managers at each of the three Trusts.

It became clear that TYPPEX features an unusual degree of complexity in terms of data selection and record linkage that was outside the scope of the basic DLPT template. The challenges stem from the limitations of the clinical database export specification when determining record eligibility.

In summary, while there are three NHS 'sites', the unit of randomisation (cluster) is a team of therapists recruited within the IAPT services of these sites. The therapist team is loosely (but not fully) aligned with existing locality teams within the service. In some cases (as represented in the diagram above), a therapist team may be made up of more than one IAPT locality

teams. Patient data required for the trial is from the caseload of participating therapists, but only where the participating therapist has completed a participant questionnaire (CAPE-P15) with the patient.

In addition to data linkage and pseudonymisation, the functionality of the TYPPEX DLPT was extended in order to complete the required processing to deal with this added complexity.

2. Data Flow Diagram (DFD)

As a starting point for designing the DFD, we considered a) established DFD design theory, b) previous examples of clinical research and trial DFDs, and c) the requirements and recommendations of the national health data custodian in England (NHS Digital).[5]

We built the DFD template for clinical trials in MS PowerPoint. The template consists of a key/objects page containing elements typically needed for this type of DFD, prompting the use of clear and helpful labels and symbols. The second page presents an example layout built for the TYPPEX WP4 trial.

The TYPPEX WP4 example incorporates many of the features found in a trial with complex data flows and can be edited and adapted to the needs of other trials and research projects.

Results and Conclusion

We selected one PCMIS site (Sussex Partnership NHS Foundation Trust) as the test site for the DLPT. The TYPPEX DLPT was sent to the site data manager alongside a request for a data cut using the conventional process (whereby the site removes identifiable fields from the exported clinical data and uploads it to a shared OneDrive folder for processing by the trial manager).

Initially, the DLPT failed due to the use of the XMATCH function within the processing code. XMATCH searches for a specified item in an array or range of cells, and then returns the item's relative position. It was first introduced to MS Excel in 2019. However, the older version of Excel used by the Trust did not support this function. The code was revised to use the older MATCH function, and the DLPT ran successfully.

The site data manager uploaded the record linked, pseudonymised .csv file to our secure OneDrive folder for checking. The file was in the format expected and could be directly imported into the REDCap trial database by the trial manager. No records were lost during this process, and all consented trial participants held in the trial database were correctly matched with clinical data from the Trust.

Following this successful test run, we plan to roll out the TYPPEX DLPT to a second site, and adapt the data specification it contains to the third site, which uses a different patient management system. The data processing done by the DLPT at site represents a significant time saving for the trial manager (10 minutes checking vs 4-5 hours of manual processing per site per month). Running the DLPT takes the site user a couple of minutes, and is therefore

comparable to the process it replaces (manually deleting and replacing identifying fields with a non-identifiable record ID).

We have demonstrated that the Data Linkage and Pseudonymisation Tool can be implemented in a cluster randomised trial with complex data linkage requirements. The DLPT has given the TYPPEX trial a quick, consistent data processing approach which reduces the burden on CTU staff (without adding to site burden) and puts the data in the required format for incorporation into the main database, and subsequent analysis. The potential future uses of the DLPT are already in discussion for grant applications in development, notably as a solution for working with research-naïve sites with limited data processing resources or skills. It should also be noted that use of the DLPT is not limited to NHS and clinical service providers. It could be used to process data from a range of sources, for example, where a trial requires information from a participant's electronic school or social care record to be linked with their research data.

We hope that researchers will use our template data flow diagram to consider the data transfers, storage and processing taking place in complex trials, and to present this information to stakeholders, ethics committees and regulatory bodies with greater clarity.

Dissemination

The DLPT, accompanying templates and guidance document, and the template data flow diagram are available via the Norwich CTU website: https://norwichctu.uea.ac.uk/DLPT/ . These resources will be advertised more widely to CTUs via the UKCRC newsletter.

We encourage the use of these resources by other researchers, and welcome feedback and suggestions for developing further functionality.

Acknowledgements

Funding acknowledgement and Disclaimer:

This project is funded by the National Institute for Health Research (NIHR) CTU Support Funding scheme. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Contribution of authors:

Polly Ashford (Norwich CTU Research Lead, Service Delivery Interventions/Routine Data) and Martin Pond (Head of Data Management) co-designed and developed the data linkage tool, data flow diagram and accompanying guidance.

Matt Hammond (Deputy Director) provided oversight of the project.

The authors would like to thank Stacey Mitchell (Sussex Partnership NHS Foundation Trust) for her patient help with testing, troubleshooting and providing feedback on the TYPPEX DLPT.

References

1. P. Bower, P. Wallace, E. Ward, J. Graffy, J. Miller, B. Delaney, A. Kinmonth, Improving recruitment to health research in primary care. *Fam Pract*., 2009;26;391.
2. M. Kho, M. Duffett, D. Willison, D. Cook, M. Brouwers, Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ* 2009; 338;822.
3. P. Ashford, C. Knight, M. Heslin, *et al.* Treating common mental disorder including psychotic experiences in the primary care improving access to psychological therapies programme (the TYPPEX study): protocol for a stepped wedge cluster randomised controlled trial with nested economic and process evaluation of a training package for therapists. *BMJ Open* 2022;12:e056355.
4. P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, J.G. Conde, Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009;42(2);377-81.; P.A. Harris, R. Taylor, B.L. Minor, V. Elliott, M. Fernandez, L. O'Neal, L. McLeod, G. Delacqua, F. Delacqua, J. Kirby, S.N. Duda, REDCap Consortium, The REDCap consortium: Building an international community of software partners. *J Biomed Inform.* 2019;95;103208.
5. NHS Digital. 2022. Data Access Request Service (DARS) guidance - NHS Digital. [online] Available at: <https://digital.nhs.uk/services/data-access-request-service-dars/dars-guidance> [Accessed 30 June 2022].

Appendices

More details about the background research, design and implementation of the tools produced in this report can be found in the accompanying document:

NIHR132544 Extended Project Report 30.06.2022

The following project output files can be found at: https://norwichctu.uea.ac.uk/DLPT/

NCTU DLPT v1.0.xlsm
NCTU DLPT Implementation Guide v1.0, 30.06.2022
Template DLPT Guide for Sites v1.0, 30.06.2022
NCTU Data Flow Diagram Template v1.0, 20.06.2022

Conflict of interest declaration

The authors had no conflicts of interest relevant to this project.

(Word count: 2441)